HISTOGRAM EQUALIZATION AND NOISE MASKING FOR ROBUST SPEECH RECOGNITION

Xueru Zhang, Kris Demuynck, Hugo Van hamme*

Katholieke Universiteit Leuven, Department of Electrical Engineering - ESAT,
Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium
{Xueru.Zhang,Kris.Demuynck,Hugo.Vanhamme}@esat.kuleuven.be

ABSTRACT

Mismatch between training and test conditions deteriorates the performance of speech recognizers. This paper investigates the combination of parametric histogram equalization (pHEQ) and noise masking to compensate for the mismatch caused by additive noise. The proposed front-end maps the distribution of the observed power spectrum vectors to a target distribution. The target distribution matches the distribution of the noise free training data except for an artificially reduced signal-to-noise ratio. Different power spectrum estimation algorithms are used to estimate the noise distribution as used internally by pHEQ more reliably under non-stationary noise conditions. The proposed front-end is evaluated on the Aurora4 database and shows a significant improvement w.r.t. mean-normalized Mel-frequency spectral coefficients. Moreover, the performance could be further improved if better estimates of the instantaneous noise power spectrum were available.

Index Terms— Histogram equalization, speech recognition, noise masking, noise power spectrum

1. INTRODUCTION

Any variability in the audio channel deteriorates the performance of a speech recognition system. This is mainly because channel variability results in linear or non-linear transformations of the speech signal [1], causing a mismatch between training and test conditions. In this paper we focus on the situation where the mismatch between training and test data is caused by additive noise.

Many methods have been proposed to improve the robustness of speech recognition systems by trying to invert the signal transformations caused by the channel differences between training and test conditions. Cepstral Mean Normalization (CMN) and Mean and Variance Normalization (MVN) are two examples. CMN removes convolutional distortions by subtracting the cepstral mean from the cepstral feature vectors [2]. MVN extends on CMN by also normalizing the variance of the acoustic feature vectors [3]. Although CMN and MVN can be used to compensate for linear transformations such as those caused by convolutional channel distortions, they are less effective when dealing with non-linear transformations resulting from the presence of for example additive noise in the channel.

Histogram equalization (HEQ) on the other hand can cope with non-linear transformations [1, 4]. The principal idea of histogram equalization is to transform the distribution of the observed acoustic feature vectors as to match a target distribution [4]. Another non-linear technique is noise masking [5, 6]. Noise masking increases the accuracy of speech recognition systems in the presence of noise by masking out low-energy events. In [6] for example, this is achieved by adding small amounts of artificial noise to the clean speech signal in order to increase the noise immunity of the system.

Techniques that cope with additive noise typically require a good estimate of the amount of noise present in the signal. Estimating the noise power spectrum from the noisy speech signal is a challenging problem, especially under non-stationary noise conditions. Improved Minima Controlled Recursive Averaging (IMCRA) as proposed by Cohen [7] is one method to estimate noise power spectrum in an adverse environment. An alternative method using a look-ahead factor was proposed by Rangachari in [8].

In this paper we propose an elegant combination of parametric histogram equalization (pHEQ) and noise masking for dealing with the mismatch between training and test conditions under additive noise. pHEQ maps the observed data distribution to a parametric target distribution, typically a mixture of two Gaussian densities. The parametric nature of the transformation makes it easy to integrate additional knowledge such as the output of a noise power spectrum estimator. Adjusting the parameters of the observation distribution in pHEQ based on Cohen's (IMCRA) or Rangachari's noise estimator makes the method more adept when dealing with non-stationary noise.

This paper is organized as follows. In section 2, we explain parametric histogram equalization. Section 3 describes the noise masking technique and how noise masking is combined with parametric histogram equalization. We illustrate the role of the noise power spectrum estimator and briefly review Cohen's (IMCRA) and Rangachari's noise power spectrum estimation algorithms in section 4. In section 5, the proposed methods are evaluated on the Aurora4 database and the experimental results are analyzed. Finally, conclusions are presented in section 6.

2. PARAMETRIC HISTOGRAM EQUALIZATION

Histogram equalization is used to reduce the mismatch between training and test conditions. HEQ maps the distribution of the observation to a target distribution. Our proposed front-end with pHEQ in an automatic speech recognition (ASR) system is shown in Fig. 1, where the "MIDA" block reduces the dimension of the feature vector parameters and decorrelates them [9]. The rest of the ASR system can be seen from [9]. There are different possible positions for pHEQ in the front-end [4]. In this paper, pHEQ is applied before the Mel-filter bank, since this configuration consistently outperformed the alternatives in our preliminary experiments. The input of the pHEQ algorithm is the logarithm[1] of the power spectrum. pHEQ

*This author is also affiliated with IBBT (Interdisciplinary Institute for Broadband Technology).

[1]Superscript "log" in the formula's represents logarithm domain.

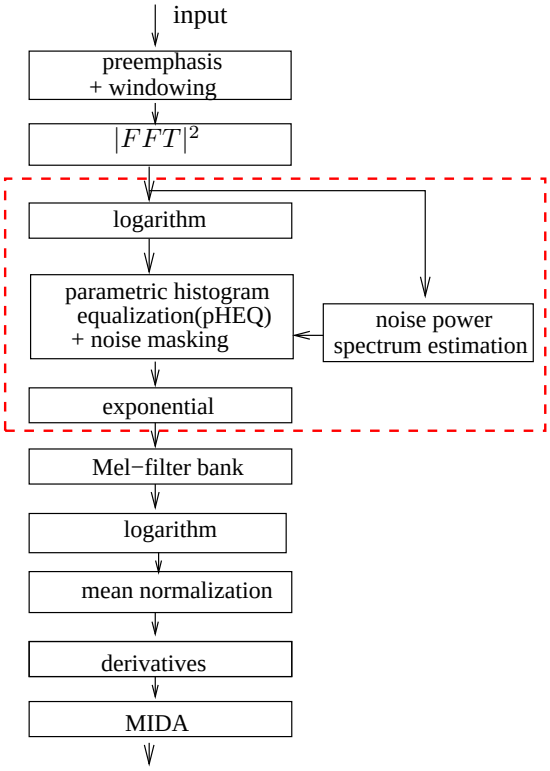


Fig. 1. The proposed speech recognition system front-end.

is applied to each frequency band independently, basically assuming little to no (usable) correlation between the noise and speech distributions of the different frequency bands.

In pHEQ, the cumulative distribution function (CDF) $C_X(P_X^{\log})$ of the observation is mapped to a target CDF $\mathcal{C}_Y(P_Y^{\log})$, with $P_X^{\log}$ and $P_Y^{\log}$ being log power spectrum values. The target distribution is typically estimated as the average distribution of the noise free training data. Both during training and testing the observed data is transformed as to match the target CDF as good as possible. The observation and target probability density functions (PDFs) $p_X(P_X^{\log})$ and $p_Y(P_Y^{\log})$ can be approximated reasonably well by a bimodal Gaussian process [4]. The bimodal Gaussian statistics form a simple Gaussian Mixture Models (GMM) for which the parameters can be efficiently estimated using Expectation Maximization (EM). Assuming a reasonable signal-to-noise ratio (SNR), the two Gaussians will correspond to the noise and the speech parts of the observation respectively. Using the symbols λ_n, $\mu_n^{\log}$, $\sigma_n^{2;\log}$ and λ_s, $\mu_s^{\log}$, $\sigma_s^{2;\log}$, for the noise and speech mixture weight, mean, and variance respectively, the parametric nature of the observation and target CDF is made explicit by eqn. (1) and eqn. (2).

$$\mathcal{C}_X(P_X^{\log}) = \mathcal{F}_X(P_X^{\log};\ \lambda_{X;n}, \mu_{X;n}^{\log}, \sigma_{X;n}^{2;\log}, \lambda_{X;s}, \mu_{X;s}^{\log}, \sigma_{X;s}^{2;\log}) \quad (1)$$

$$\mathcal{C}_Y(P_Y^{\log}) = \mathcal{F}_Y(P_Y^{\log};\ \lambda_{Y;n}, \mu_{Y;n}^{\log}, \sigma_{Y;n}^{2;\log}, \lambda_{Y;s}, \mu_{Y;s}^{\log}, \sigma_{Y;s}^{2;\log}) \quad (2)$$

Considering that the size of the observed data set is far smaller than the size of the target data set, the estimated variance $\sigma_{Y;n}^{2;\log}, \sigma_{Y;s}^{2;\log}$ from the target will be more reliable than the estimated variance $\sigma_{X;n}^{2;\log}, \sigma_{X;s}^{2;\log}$ from the observation. Furthermore, our preliminary experiments show that any scaling of the logarithm spectrum due to different variances, which generates strong non-linear transformations in the power domain, deteriorates the results substantially. By

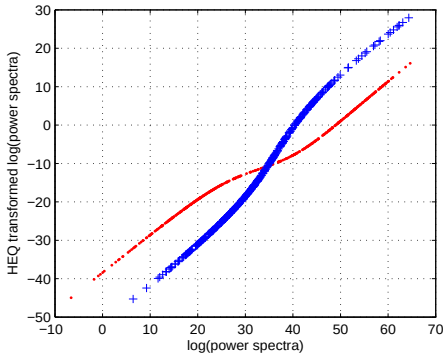


Fig. 2. Parametric histogram equalization transformation curves. The abscissa is the logarithm power spectrum $P_X^{\log}$ of one observation. Ordinate is the corresponding pHEQ transformed logarithm power spectrum $P_Y^{\log}$. The curve represented by plus sign shows the case when the observation SNR $<$ the target SNR. The dotted line describes the case when the observation SNR $>$ the target SNR.

replacing the observed variances by the target variances, both ends of the pHEQ mapping curve have a slope of 1. The weight and mean values for the observation are estimated by EM. In other words, for our experiments eqn. (1) was replaced by eqn. (3).

$$\mathcal{C}_X(P_X^{\log}) = \mathcal{F}_X(P_X^{\log};\ \lambda_{X;n}, \mu_{X;n}^{\log}, \sigma_{Y;n}^{2;\log}, \lambda_{X;s}, \mu_{X;s}^{\log}, \sigma_{Y;s}^{2;\log}) \quad (3)$$

Given a sequence of observations $P_X^{\log}$ with a CDF as given in eqn. (3), pHEQ is to find the corresponding sequence of $P_Y^{\log} = \text{pHEQ}(P_X^{\log})$ values that match the target CDF as given in eqn. (2). From the required equality of the CDF's as expressed in eqn. (4), one can easily derive the transformation function as given by eqn. (5).

$$\mathcal{C}_Y(P_Y^{\log}) = \mathcal{C}_X(P_X^{\log}) \quad (4)$$

$$P_Y^{\log} = \text{pHEQ}(P_X^{\log}) = \mathcal{C}_Y^{-1}(\mathcal{C}_X(P_X^{\log})) \quad (5)$$

Fig. 2 shows two examples of pHEQ transformation curves, with different observation SNR and different target SNR. The plus sign transfer function increases the observation SNR, similar to spectral subtraction. The dotted transfer function decreases the observation SNR, similar to noise masking (see section 3).

3. NOISE MASKING

Noise masking improves the speech recognizer performance by reducing the signal-to-noise ratio to a fixed value. In [5], a noise masking value substitutes the output of the filter bank if the output falls below the masking value. In [6], noise masking is implemented by adding extra artificial noise to the speech signal in order to attain the desired SNR. Noise masking removes low-energy spectral details that are only visible in (very) clean speech conditions but which are irrelevant in more realistic, i.e. noisy, conditions. By this way, the acoustic features learned under "clean" conditions will be more similar to the acoustic features one can expect in noisy conditions [6], without losing much relevant speech details.

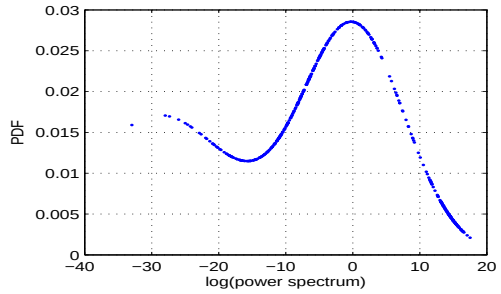
Considering that both pHEQ and noise masking help in decreasing the variability of the speech signal, we propose to combine both methods to improve the performance of speech recognition systems. Noise masking is combined with pHEQ by lowering the SNR $\mu_{Y;s}^{\log} - \mu_{Y;n}^{\log}$ of the target distribution in pHEQ to a fixed value. This new

target SNR is chosen to decrease the mismatch between training and test data and at the same time keep most of the speech information. Therefore, the target CDF in pHEQ eqn. (2) is rewritten as eqn. (6).

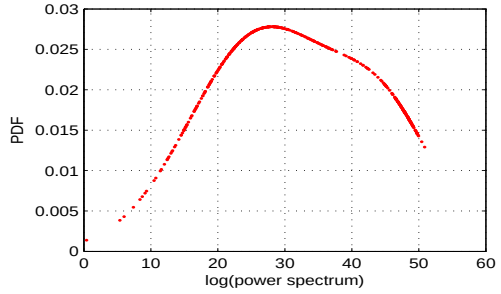
$$\mathcal{C}_Y^M(P_Y^{\log}) = \mathcal{F}_Y^M(P_Y^{\log}; \lambda_{Y;n}, \mu_{Y;n}^{M;\log}, \sigma_{Y;n}^{2;\log}, \lambda_{Y;s}, \mu_{Y;s}^{M;\log}, \sigma_{Y;s}^{2;\log}) \quad (6)$$

where $\mu_{Y;n}^{M;\log}, \mu_{Y;s}^{M;\log}$ are the new noise mean and new speech mean.

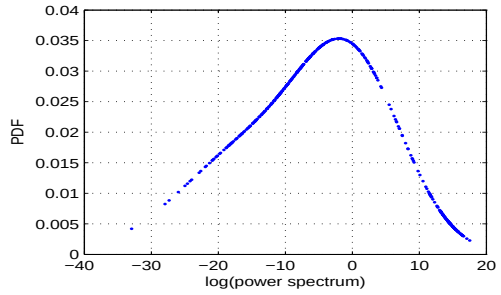
Fig. 3 shows an example of a clean speech target PDF, one observation PDF for noisy speech, and the adjusted target PDF taking into account noise masking. The dissimilarity between Fig. 3(a) and 3(b) illustrates the mismatch caused by additive noise. Adjusting the target PDF reduces the dissimilarity (Fig. 3(b) versus 3(c)).



(a) Target PDF



(b) Observation PDF



(c) New adjusted PDF based on eqn. (6).

Fig. 3. (a) An example of a target PDF without noise masking, with $\mu_{Y;n}^{\log} = -28.9$ dB, $\mu_{Y;s}^{\log} = 0$ dB. (b) An observation PDF example. (c) The adjusted target PDF, with $\mu_{Y;n}^{M;\log} = -15$ dB, $\mu_{Y;s}^{M;\log} = 0$ dB.

4. NOISE POWER SPECTRUM ESTIMATION

The bimodal Gaussian process estimates the observation noise mean $\mu_{X;n}^{\log}$ under the assumption that the noise is stationary. However, in reality, most environments exhibit moderate to high non-stationarity. Hence, a more accurate estimation of the noise statistic $\mu_{X;n}^{\log}$ should result in better CDF mapping in pHEQ. Cohen [7] and Rangachari [8] have proposed different techniques to estimate non-stationary noise under adverse environments. Both these methods estimate the noise power spectrum by averaging past power spectrum values P_X using a time-varying frequency-dependent smoothing parameter. The smoothing parameter is updated by the speech presence probability $p(t,k)$ in each frame and subband.

4.1. Improved Minima Controlled Recursive Averaging

The IMCRA [7] noise estimation process is defined by eqn. (7) and eqn. (8).

$$\widehat{\mu}_{X;n}(t+1,k) = \tilde{\alpha}_d(t,k)\widehat{\mu}_{X;n}(t,k) + (1-\tilde{\alpha}_d(t,k))P_X(t,k) \quad (7)$$

$$\tilde{\alpha}_d(t,k) = \alpha_d + (1-\alpha_d)p(t,k) \quad (8)$$

with α_d a constant. The speech presence probability is a function of the a priori probability for speech absence and the a priori SNR. The a priori probability is controlled by the minima values of the smoothed noisy power spectrum. There are two iterations of smoothing and minimum tracking to estimate the a priori probability. Both iterations are carried out in time and frequency. Therefore, the correlation of speech presence in neighboring subbands of continuous frames is taken into account. The first iteration provides a rough voice activity detection in each subband. The second iteration provides robust minimum tracking by excluding relatively strong speech components. A bias factor β is introduced in IMCRA to compensate for the bias of the noise power spectra toward lower values. The updated formula for eqn. (7) is given by eqn. (9).

$$\mu_{X;n}(t+1,k) = \beta\widehat{\mu}_{X;n}(t+1,k) \quad (9)$$

4.2. The Rangachari noise estimation algorithm

The smoothing parameter in Rangachari's noise estimator [8] is adjusted as given in eqn. (8). However, the estimation of the speech presence probability $p(t,k)$ is different from IMCRA. Instead of finding the minimum of the noisy speech within a certain window, Rangachari's algorithm tracks the minimum by averaging the past smoothed noisy power spectrum $P_X^S(t,k)$. The minimum $P_{X;\min}^S(t,k)$ is given as eqn. (10).

$$\begin{aligned} \text{If} \quad P_{X;\min}^S(t-1,k) < \; & P_X^S(t,k), \text{ then} \\ P_{X;\min}^S(t,k) = \; & \gamma P_{X;\min}^S(t-1,k) \\ + \; & \tfrac{1-\gamma}{1-\zeta}(P_X^S(t,k) - \zeta P_X^S(t-1,k)) \\ \text{else} \\ P_{X;\min}^S(t,k) = \; & P_X^S(t,k) \end{aligned} \quad (10)$$

with γ and ζ constants. The rough speech presence decision $I(t,k)$ is given by comparing the ratio of the smoothed noisy power spectrum and its local minimum with a frequency-dependent threshold. The speech presence probability $p(t,k)$ is given by eqn. (11).

$$p(t,k) = \alpha_p p(t-1,k) + (1-\alpha_p)I(t,k) \quad (11)$$

with α_p a constant.

5. EXPERIMENTS

5.1. Database

The performance of the proposed compensation algorithm is evaluated on the Wall Street Journal (WSJ0) based Aurora4 database. For our experiments, we use the clean condition training set and test sets 01-07. Test set 01 contains noise free data. Test sets 02-07 were

Test	base	AFE	IMCRA	Rang	Instan	Instan2
01	5.81	6.20	6.86	6.16	6.86	5.72
02	15.28	14.38	10.57	10.67	10.39	8.87
03	31.87	23.07	25.71	22.85	11.71	16.59
04	40.76	30.17	31.40	30.64	12.98	20.40
05	34.04	26.45	26.68	26.45	13.94	21.39
06	27.93	24.92	25.33	21.86	10.14	14.22
07	36.52	23.89	26.28	26.42	14.50	20.94
Avg	27.46	21.30	21.83	20.72	11.50	15.45
Imp		17.20	16.29	21.28	46.91	38.62

Table 1. WER in "%" on the Aurora4 test sets under clean condition training. *base* is the baseline system without compensation. *AFE* is the ETSI ES 202 050 Advanced Front-end. *IMCRA* and *Rang* refer to the proposed front-end with the IMCRA noise estimator and Rangachari's noise estimator. *Instan* is the proposed front-end with known instantaneous noise power spectrum. *Instan2* is the same as *Instan*, except that the derivatives are calculated before the pHEQ transform. *Avg* is the average WER over all test sets. *Imp* is the average relative improvement compared to the baseline.

created by artificially adding noise to the clean data from test set 01. The noise types used are: set 02 (car), set 03 (babble), set 04 (restaurant), set 05 (street), set 06 (airport), and set 07 (train).

5.2. Results

In this section, we evaluate the performance of the proposed front-end and an uncompensated front-end (baseline system). The latter is the front-end without the pHEQ, noise masking and noise power spectrum estimation techniques. The results for the European Telecommunications Standards Institute (ETSI) distributed speech recognition (DSR) ES 202 050 Advanced Front-end (AFE) with frame dropping [10] are also given. The performance of the proposed front-end is evaluated with different noise estimators. The proposed compensation algorithm is applied to each utterance of the training and test data with $\mu_{Y;n}^{M;\log} = -15$ dB, $\mu_{Y;s}^{M;\log} = 0$ dB.

Table 1 lists the word error rate (WER) and the average relative improvement for different front-ends and noise types. The *Instan* and *Instan2* columns list the results of the proposed front-end with known instantaneous noise power spectrum $\mu_{X;n}^{\log}$ for test sets 02-07. The noise power spectrum values for the clean data (training data and test set 01) for *Instan* and *Instan2* were estimated using IMCRA. The first and second order time derivatives are estimated after pHEQ except for *Instan2* which uses the baseline, i.e. uncompensated, front-end derivatives.

All noise robust front-ends clearly improve over the baseline. *AFE* and *IMCRA* give similar performance while *Rang* show a somewhat larger relative improvement. Unlike the IMCRA noise estimator, Rangachari's noise estimator does not depend on some fixed window length and hence Rangachari's estimated noise power spectrum can be updated faster under non-stationary noisy conditions. The importance of closely tracking the noise is further illustrated by the *Instan* and *Instan2* results. Both yield substantially higher relative improvements than any of the other methods. This implies that improvements in estimating the instantaneous noise power spectrum will automatically lead to a further decrease in WER when using the proposed front-end. Comparing *Instan2* with *Instan* shows the effect of a commonly used alternative when modifying the features: only modify the static features and keep the original dynamic features. This improves the clean speech result at the cost of the noisy speech results. The same effect was observed when altering the *IMCRA* or *Rang* setup, although somewhat less pronounced.

6. CONCLUSIONS

In this paper, we have presented a front-end that combines parametric histogram equalization and noise masking to reduce the mismatch between training and test conditions. Parametric histogram equalization maps the observation cumulative density function to a target cumulative density function. Noise masking increases the immunity of the speech recognition system by lowering the signal-to-noise ratio to a fixed value. A noise power spectrum estimator is combined with our parametric histogram equalization and noise masking to improve the system's performance when dealing with non-stationary noise. Both Cohen's Improved Minima Controlled Recursive Averaging method and Rangachari's noise power spectrum estimation were evaluated. The proposed front-end reduces the word error rate of the speech recognition system w.r.t. our baseline system and matches the performance obtained when using the AFE. Analysis of the proposed front-end under the assumption that the instantaneous noise power spectrum is known, shows that with the arrival of better instantaneous noise power spectrum estimators the word error rate will further reduce.

7. REFERENCES

[1] Ángel de la Torre, José C. Segura, Carmen Benítez, Antonio M. Peinado, and Antonio J. Rubio, "Non-linear transformations of the feature space for robust speech recognition," in *Proceedings of ICASSP 2002*, 2002, pp. 401–404.

[2] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, 2001.

[3] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," in *Speech Communication*, 1998, vol. 25, pp. 133–147.

[4] Sirko Molau, Michael Pitz, and Hermann Ney, "Histogram based normalization in the acoustic feature space," in *Proc.ASRU2001 - Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, 2001.

[5] D. Klatt, "A digital filter bank for spectral matching," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP*, 1976, vol. 1, pp. 573–576.

[6] D. Van Compernolle, "Noise adaptation in a hidden markov model speech recognition system," in *Computer Speech Language*, 1989, vol. 3, pp. 151–167.

[7] Israel Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," in *IEEE Transactions on Speech and Audio Processing*, 2003, vol. 11, pp. 466–475.

[8] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," in *Speech Commun.*, 2006, vol. 48, pp. 220–231.

[9] K. Demuynck, *Extracting, Modelling and Combining Information in Speech Recognition*, Ph.D. thesis, K.U.Leuven, 2001.

[10] ETSI standard doc., "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.